Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2019;5(3): 183-193

Original Article

A New Application of Louvain Algorithm for Identifying Disease Fields Using Big Data Techniques

Saeed Shirazi[1], Hamed Baziyad[1], Naser Ahmadi[2], Amir Albadvi[3]*

[1]Department of Information Technology, Faculty of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran.
[2]Department of Biostatistics, Faculty of Paramedical Science, Shahid Beheshti University of Medical Science, Tehran, Iran.
[3]Department of Information Technology, Faculty of Industrial and Systems Engineering, Tehran, Iran.

ARTICLE INFO

ABSTRACT

Received 17.01.2019
Revised 01.05.2018
Accepted 13.06.2019
Published 01.08.2019

Key words:
Big data;
Graph theory;
Community detection;
Drug prescription

Background and aim: Recently, the use of data science techniques in healthcare has been increased remarkably. Community detection as one the important methods of data science is utilized in the health domain.
Methods: This paper detects disease areas based on combination of big data and graph mining methods on drug prescriptions. At first, network of prescription is designed, and Louvain algorithm is applied for community detection of 50000 Iranian prescriptions in 2014 gathered from the Iranian Health Insurance Organization. We use modularity metric for validation of the results and the experts' opinion as the external validation of communities.
Results: The outputs are consist of six communities. These communities are labeled based on experts' opinion that present the disease fields.
Conclusion: The Louvain algorithm has the ability to detect the major communities of the prescription database with an acceptable accuracy. We have proven that these communities present the disease fields.

Introduction

Big data is one of the important techniques in the area of healthcare, which has a growing presence in healthcare innovation.[1] Many healthcare and medical operations can be supported by big data. Controlling diseases and managing health are two related operations.[2] The volume of data generated in the field of health informatics has increased considerably and analyzing such huge data volume creates new opportunities for knowledge discovery.[3] Source of big data in healthcare comes from four primary categories, namely Electronic Health Record (EHR), Medical Imaging Data, Unstructured Clinical Notes, and Genetic Data.[4] EHR can be divide into some classes such as laboratory results, billing data, medication records, and test details.[5] In this paper, data lies into EHR category.

Besides, Community detection plays an essential role in many fields such as sociology, biology and computer science.[6] A community is regarded as a dense set of connected nodes, which are linked to other parts of graph sparsely.[7] In other words, a community is comprised of groups of nodes which have more relations with each other rather external groups.[8] The measure of modularity is one of the best criteria for evaluation of the communities' quality.[9] Discovering high-quality communities in the large scale

* . Corresponding Author: albadvi@modares.ac.ir

Please cite this article in press as Shirazi S, Baziyad H, Ahmadi N, Albadvi A. A New Application of Louvain Algorithm for Identifying Disease Fields Using Big Data Techniques. J Biostat Epidemiol. *2019; 5(3): 183-193*

networks is regarded as a challenge in data mining approach.[10]

In this study, community detection approach is applied in an innovative way of finding similar groups of prescriptions by using the co-occurrence of prescriptions technique. One of the reasons for applying community detection in this paper is lack of informative features in prescription data set. This matter prevents us from using well-knows data mining methods. Use of parallel computing provided appropriate situations for dealing with large-scalability and complexity properties of the network.

Furthermore, Louvain algorithm is implemented for finding the communities in this research owing to its acceptable time complexity of $O(n \log n)$. This provides an opportunity to analyze a sizeable amount of data for reaching more acceptable results. The Louvain algorithm proposed by Blondel et al. aims at maximizing the modularity metric. At first, this algorithm assumes that each node represents a community by itself. Afterward, iteratively, it considers every node with its adjacent nodes as a community. If the modularity increases by this consideration, it merges the node with its nearby nodes and continues this operation until no further improvement in modularity can happen.[11]

Related Works

Some papers in community detection of large scale/ big data networks and complex networks have focused on methods and algorithms.

It is believed that dealing with large-scale graphs with billions of edges and nodes is changed into a challenging issue for big data researchers. Dabas et al. introduced hybrid method named random walk fast greedy for evaluation of large-scale graph in the field of community detection.[12] The algorithm has been improved significantly compared with some existing algorithms such as betweenness, walk trap, leading eigenvector, and infomap. Finding community structure has been growing fast recently. As a central defect in existing algorithms of community detection, they ignore information topology. Chopade and Zhan suggested a more accurate method regarding information topology for community detection in large scale/big data networks.[13] The proposed algorithm recovers communities successfully in real networks. Moreover, Chopade and Zhan introduced a novel game-theory approach for community detection in large-scale complex networks, which is applicable to networks. The scalability of this method enables it to partition a graph into dense communities in big data.[14] Recently, an investigation in the field of overlapping community detection has grown tides of attentions. Overlapping community detection tries to discover communities of specific nodes, which are laid in multiple communities. Overlapping community detection faces with some limitations such as high computational complexity, high overlapped communities, low steadiness (diversity of solutions during the time) and unidentified nodes (disability to lie every node into communities). To solve such problems a new algorithm was proposed by Long. Experimental outputs indicate efficiency and accuracy of the proposed algorithm.[15] Other studies have been investigated in the area of introducing algorithms for large-scale community detections.[16–19] In this paper, in order to deal with complex and large-scale network, The Louvain algorithm[20] is utilized in a parallel way which reduces computation time efficiently.

In the field of large network, network science, big data techniques such as Spark[21,22] and Hadoop[23–25], were used extensively. Network science of large-scale networks has been investigated in many fields such as social networks[26], airline networks[21], anti-money laundering[41], real-world mobile phone data[27,] and health[28]. For the first time, big data was introduced by Michael Cox and David Ellsworth about data visualization challenge in computer systems in 1997.[29] Beginning of 2009 is known as a revolutionary phase of big data analysis.[30] Big data is explained by three main parameters, namely the amount of data (volume), type of data (variety) and the rate of data generation (velocity).[31] Big data analysis provides novel insights in the health field for stakeholders such as advance personalized care, improving patient outcomes and avoid unessential costs.[32] Big data usage has cut $300 million annually in the healthcare industry in the United States.[33] Big data analytic methods have been used in many fields of health such as health services,[34] biological discovery,[35] medicine,[36–38] cancer,[39,40] drug discovery[41], and many other fields. In addition to these areas, many methods and algorithms have been used in health Using big data technics and tools aim to gain more accurate results and with acceptable runtime complexity. This paper focuses on network science methods, especially community detection for knowledge discovery about relationships of prescriptions using big data techniques.

Investigation of drug-drug interactions has been changed into a novel discovery tool, which can reveal drug action patterns. Udrescu et al. presented a graph-based approach for analyzing drug-drug interactions.[42] Designed network included of 1141 nodes indicating drugs and interactions between drugs referring to the edges. The network comprised of 1688 edges. The community detection method based on Girvan and Newman's algorithm was conducted, ending up to nine communities. As a result, by the use of community detection in drug-drug network, drugs categories and relationships between drugs are disclosed. Most genes, proteins and other ingredients, conduct their duties in complex network actions and reactions. In order to discover biological information from interaction network, Chautard et al. presented a topological analysis based on protein-protein interaction network. The introduced method revealed some knowledge about functions of protein, metabolic signaling pathways, physiological procedures, the molecular foundation of some illnesses such as cancer and infectious diseases.[43]

The combination of big data techniques and network science provides many opportunities for knowledge discovery. However, studies about the combination of big data and network science, especially community detection are in their early stages. The combination of big data and community detection has been studied in areas such as big social networks, airline networks, anti-money laundering, IoT and mobile phone data. In this study, a community detection based on big data has been utilized to knowledge discovery about the prescription.

Methodology

Knowledge discovery of medical prescriptions plays a vital role in the field of healthcare. Detection of disease areas is one of the exciting researches helping data scientists for analyzing prescriptions. The main goal of this paper is detection of disease areas by applying network science methods on Iranian prescriptions in 2014 gathered from the Iranian Health Insurance Organization. Use of community detection in an innovative way is the base of the research. Indeed, in order to deal with leakage of informative fields and network sparsity, network science methods have been applied rather than standard data mining techniques.

Data Selection

At first, 50000 prescriptions were selected. Every row of the data represents a prescription consists of the prescription's id, patient's id, patient's sex, patient's age, drugs' ids, and drugs names in every prescription. We start with selecting proper features of data for further investigations. Features we need for this study are prescription id, drugs' ids, and drug names.

Preprocess

In the next step, we clean data from abnormal cells, meaning that if we have some extra characters in drug names, we remove those characters to get the data clean. For example, some cells of data have characters like "$" or "#" and we try to remove these and fetch informative data. The same thing happens for numeric values of prescription's id and drug ids. To have reliable results, we use a strategy for null values. If prescription id is missing, as the primary key of data, we remove that specific prescription from data. In case of missing drug id in a prescription, we remove that drug from prescription and at last, if some drug names are missing, we try to refill drug name using drugs dictionary, obtained from the Food and Drug Administration (FDA), based on the presence of that drug's id.

By elaborating on drug dictionary, it has been revealed that in some cases, for some drugs with the same kind, there are two or more drug ids assigned to each of them. For instance, for "ACETAMINOPHEN CODEINE (500+8)" and "ACETAMINOPHEN CODEINE 300mg-20mg" there were two different drug IDs assigned. Considering the fact that for a faster run time we do our processing only on drug IDs as a Long Scala type which is faster in processing in comparison to drug names as String data format and for final visualization and analysis we assign drug names to drug IDs using a drug dictionary. Therefore, we aimed at considering same drug types as one single drug. It goes without saying that in these cases, medications with the same type but different dosage or different way of use have different usage in the case of the patient's situation. Nevertheless, in this study considering these cases as same helps the final result.

Community detection

To analyze the data collected, we create a graph based on common medicines in each prescription. Every prescription is considered as a node in our network, and for every common medicine in two prescriptions, we create an edge between these two prescriptions using Spark framework functions mostly map, flatmap and reduce by help of the Scala programming language. By means of this method, the weight of every edge is the number of common medicines between two specific nodes. It is expected that each community detected, indicate groups of medicines presenting the particular disease area. Due to the large-scale nature of the graph, the graph was created in the Apache Spark environment, which uses parallel computing. At this step, the output graph had 50000 nodes and more than 200 million edges.

At the next step, we pass the output network to community detection algorithm aiming at finding prescriptions' communities. In this study, we use the Louvain algorithm for detecting community structures of prescriptions. Blondel et al. introduced this popular greedy algorithm for community detection–the Louvain algorithm[11], for the general case of weighted graphs. The Louvain algorithm starts by putting all vertices of a graph in distinct communities, one per vertex. It then sequentially sweeps over all vertices in the inner loop. For each vertex (i), the algorithm performs two calculations: (1) compute the modularity gain ΔQ when putting a vertex (i) in the community of any neighbor (j); (2) pick the neighbor (j) that yields the largest gain in ΔQ and join the corresponding community. This loop continues until no movement yields again. At the end of this phase, the Louvain algorithm obtains the first-level partitions. In the second step, these partitions become super vertices, and the algorithm reconstructs the graph by calculating the weight of edges between all super vertices. Two super vertices are connected if there is at least one edge between vertices of the corresponding partitions, in which case the weight of the edge between the two super vertices is the sum of the weights from all edges between their corresponding partitions at the lower level. These two steps of the algorithm are then repeated, yielding new hierarchical levels and super graphs. The algorithm stops when communities become stable and modularity metric gain no improvement. The Louvain algorithm typically converges very quickly, and it can identify communities in just a few iterations.

For visualization of resulted communities, Gephi 9.0.2 is used. Every community is represented as a single graph. Node sizes are based on betweenness centrality. Betweenness centrality is generally regarded as a measure of others' dependence on a given node, and therefore as a measure of the potential control of nodes,[44] and for ease of experts labeling process, we execute Louvain algorithm another time on communities. As a result, nodes' color in every community presents a sub-community of nodes in that specific community. Figures 1-6 illustrate communities of disease and labeling of each created community is conducted based on experts' opinion.

After running Louvain algorithm on the graph at first place, in experts point of view, there were some non-specialized drugs than had a significant impact on graph's density and as a result, of community detection output. These drugs were widely used in prescriptions and with specialized drugs. However, they did not represent any specific type of disease. For instance, painkillers and antibiotics were used in large numbers. Moreover, considering the fact that non-specialized drugs add no information about specific diseases, we aimed at removing these drugs. In addition, nodes with high degree, called 'hub' in network science, are responsible for high density of graphs. Removing these nodes reduces graph density, and resultantly, this matter reduces number of edges between communities and makes the process of finding the community structures more accurate. Based on data gathered from Drug and Food Organization, which specifies all specialized and non-specialized drugs, we dismiss non-specialized drugs from every prescription and saved only 688 drug types. This leads our study to a better result owing to wide range of prescriptions includes lots of non-specialized drugs, which are existed in many prescriptions. In addition, removing these drugs resulted in less graph density and higher modularity.

At the end of increasing modularity section, only 20441 of 50000 prescriptions (nodes) with at least one specialized drug and 45078182 common drugs in prescriptions (edges) remained. These steps resulted in a modularity measure of 0.53 and 57 communities were detected. However, only six communities consisted of more than 1000 nodes, and number of nodes in other communities was less than 100. Therefore, for further investigations, we consider small communities as anomaly and focus on six major ones.

Evaluation

At the stage of community detection, after each time running this algorithm on data, we analyze if the result of modularity metric is acceptable, means that communities are detected well or not. Furthermore, at each iteration, we visualize communities and experts decide if prescriptions in every community are related enough to each other that they can put a specific label on every community or not. At each step, if results are not proper enough, we manipulate input data of Louvain algorithm to reach outputs that are more acceptable. Thus, based on our case, we try to reduce graph density by removing uninformative data. This can help the process of finding the communities, by removing some of the edges between these communities. Eliminating non-specialized medicines is an example of manipulating input data for Louvain algorithm.

In contrary to the default execution of the Louvain, which is run sequentially, in this study the community detection algorithm was executed in a parallel way in Spark environment. Therefore, we did not expect high modularity results in comparison to sequential implementations of Louvain algorithm. However, in experts' point of view, these communities were still interlinked together and they could not put specific labels on them. Therefore, in following section we discuss how we changed the input graph to achieve higher modularity and consequently more accurate communities. Stages of the research are shown in table1.

Table 1 methodology requirements for community detection of prescriptions

Tools	**Algorithm**	**Used methods**	**Stages**	
spark framework/ scala 2.12.4/ scala IDE 4.5.0/ sbt	-	Map\Reduce	Data Selection	
spark framework/ scala 2.12.4/ scala IDE 4.5.0/ sbt	-	Map\Reduce\Filter	Preprocess	
Spark GraphX/ scala IDE 4.5.0/ scala 2.12.4	Louvain	Community detection	Analysis	Network Science
Gephi 0.9.2	Force Atlas	Betweenness centrality	Visualization	
Spark Graph/ scala 2.12.4/ scala IDE 4.5.0	Louvain	modularity metric	Modularity	Evaluation
-	-	-	Experts	

Results

Finally, six main communities were determined as a result of aforementioned methodology. Features of each community are shown in table 2. Column '**number of prescriptions**' indicates the number of prescriptions in each community and column '**Density'** means the amount of internal correlation of each community. Indeed, clusters with high density represent well development and maturity.

'Respiratory and urinary tract infections' with 6706 (35.16% of prescriptions) prescriptions has the most number of prescriptions 'Infectious diseases and pain' with 3858 (20.23% of prescriptions), 'Anaerobic infections, gastrointestinal infections and infections of women' with 3603 (18.89%), 'allergic diseases and asthma' with 1273 (6.67%), 'Depression' with 1160 (6.08%) and 'upper respiratory tract infections' with 1137 (5.96%) have included other common prescriptions, respectively. Besides, figure 1 up to figure 6, indicated related graphs. 'Allergic diseases and asthma' with density 0.444 has the most developed community. After that, 'Upper respiratory tract infectious' with 0.398, 'Depression' with 0.283, 'Anaerobic infections, gastrointestinal infections and infections of women' with 0.073, 'Infectious diseases and pain' with 0.063 and 'Respiratory and urinary tract infections' with 0.022 are laid in other places, respectively.

Table 2 Features of each community

Row	Community label	number of prescriptions	Density
1	Infectious diseases and pain	3858	0.063
2	Allergic diseases and asthma	1273	0.444
3	Depression	1160	0.283
4	Anaerobic infections, gastrointestinal infections, and infections of women	3603	0.073
5	Upper respiratory tract infectious	1137	0.398
6	Respiratory and urinary tract infections	6706	0.022
7	Total	19072	-

This study focused on finding different communities from Iranian prescriptions in 2014 gathered from the Iranian Health Insurance Organization. Fifty thousand prescriptions were analyzed based on big data and network science. The prescription-prescription network produced for community detection had an enormous amount of edges. As a matter of fact, one of the main restrictions on this paper was hardware limitations. This limitation for analyzing network with more than 200 millions of edges made us select only 50 thousand prescriptions for creating the prescription-prescription network. Moreover, using Louvain as a fast algorithm in spark framework with parallel computations made it possible to analyze a proper amount of prescriptions leading to more accurate results. Resultantly, six main communities were disclosed which have the most usage among the Iranian people, respectively.

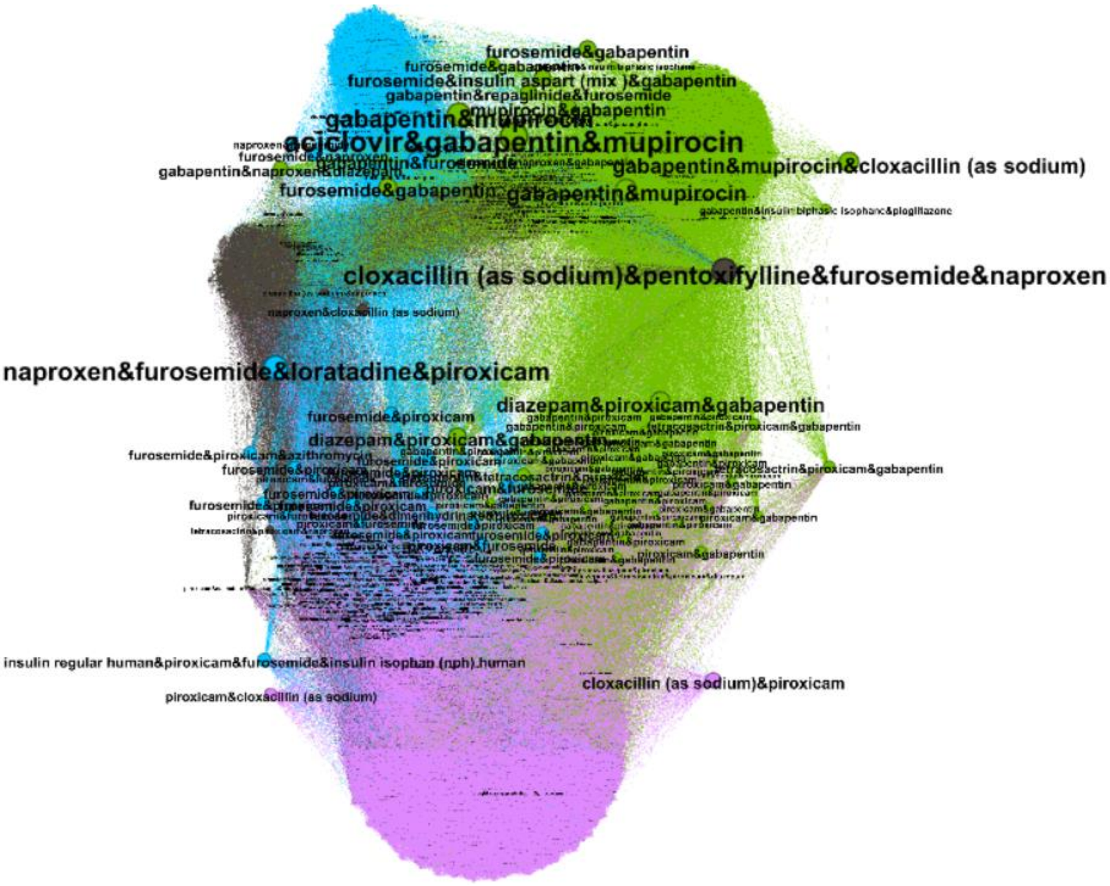
furosemide&gabapentin
furosemide&insulin aspart (mix)&gabapentin
gabapentin&repaglinide&furosemide
gabapentin&mupirocin
aciclovir&gabapentin&mupirocin
gabapentin&mupirocin&cloxacillin (as sodium)
gabapentin&naproxen&diazepam
furosemide&gabapentin
gabapentin&mupirocin
cloxacillin (as sodium)&pentoxifylline&furosemide&naproxen
naproxen&cloxacillin (as sodium)
naproxen&furosemide&loratadine&piroxicam
diazepam&piroxicam&gabapentin
furosemide&piroxicam
diazepam&piroxicam&gabapentin
furosemide&piroxicam&azithromycin
insulin regular human&piroxicam&furosemide&insulin isophan (nph) human
cloxacillin (as sodium)&piroxicam
piroxicam&cloxacillin (as sodium)

Figure.1 The ‘Infectious diseases and pain’ community

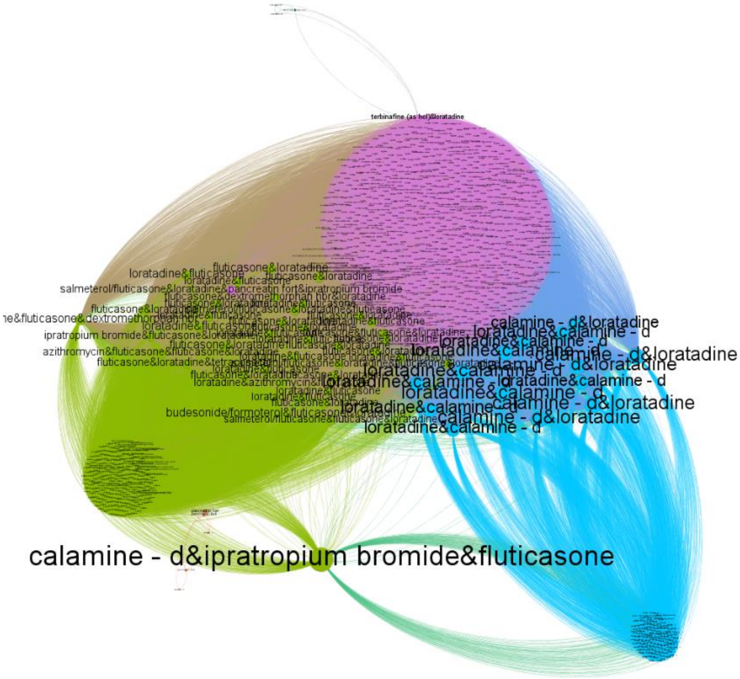
terbinafine (as hcl)&loratadine
loratadine&fluticasone
fluticasone&loratadine
salmeterol/fluticasone&loratadine&pancreatin fort&ipratropium bromide
calamine - d&loratadine
loratadine&calamine - d
calamine - d&loratadine
budesonide/formoterol&fluticasone
calamine - d&ipratropium bromide&fluticasone

Figure.2 The ‘Allergic diseases and asthma’ community

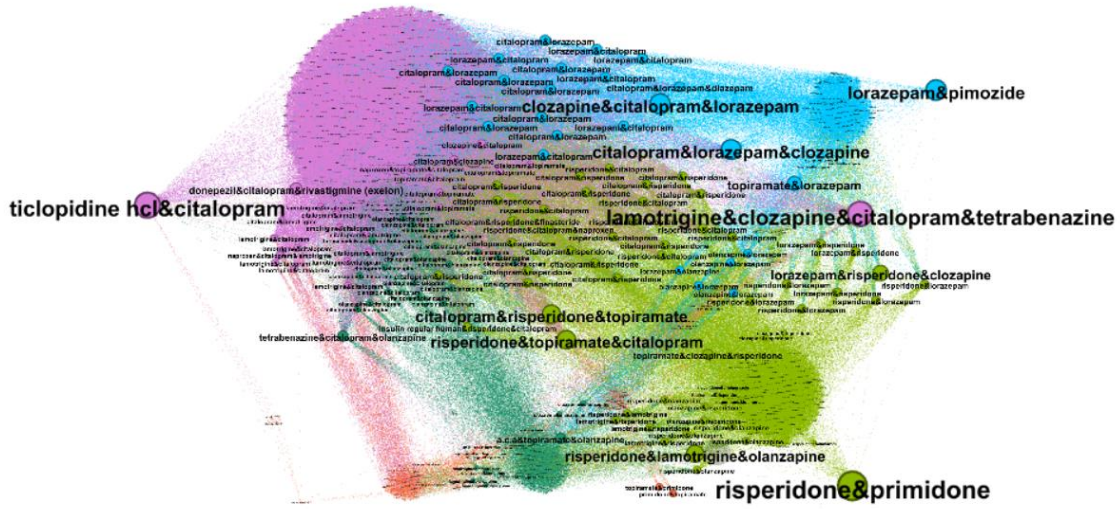
lorazepam&pimozide
clozapine&citalopram&lorazepam
citalopram&lorazepam&clozapine
topiramate&lorazepam
ticlopidine hcl&citalopram
donepezil&citalopram&rivastigmine (exelon)
lamotrigine&clozapine&citalopram&tetrabenazine
lorazepam&risperidone&clozapine
citalopram&risperidone&topiramate
risperidone&topiramate&citalopram
tetrabenazine&citalopram&olanzapine
topiramate&clozapine&risperidone
risperidone&lamotrigine&olanzapine
risperidone&primidone

Figure.3 The ‘Depression’ community

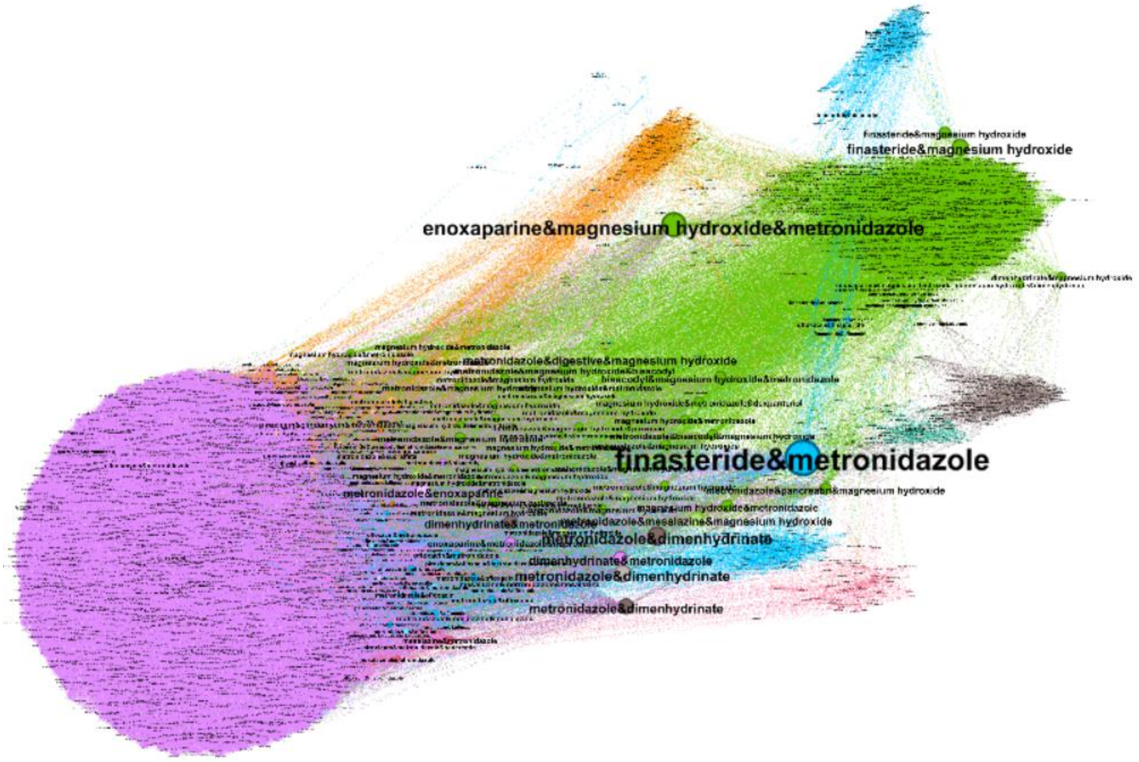
finasteride&magnesium hydroxide
finasteride&magnesium hydroxide
enoxaparine&magnesium hydroxide&metronidazole
metronidazole&digestive&magnesium hydroxide
finasteride&metronidazole
metronidazole&pancreatin&magnesium hydroxide
metronidazole&enoxaparine
magnesium hydroxide&metronidazole
metronidazole&mesalazine&magnesium hydroxide
dimenhydrinate&metronidazole
metronidazole&dimenhydrinate
dimenhydrinate&metronidazole
metronidazole&dimenhydrinate
metronidazole&dimenhydrinate

Figure.4 The ‘Anaerobic infections, gastrointestinal infections and infections of women’ community

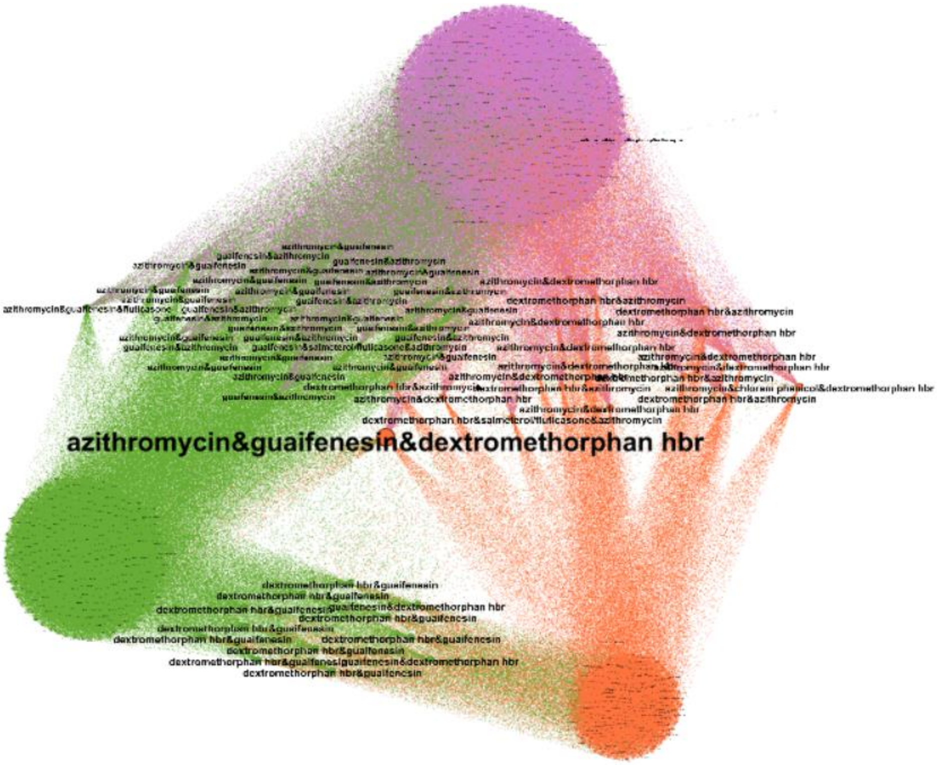
azithromycin&guaifenesin&dextromethorphan hbr
azithromycin&dextromethorphan hbr
dextromethorphan hbr&azithromycin
guaifenesin&azithromycin
azithromycin&guaifenesin
dextromethorphan hbr&salmeterol/fluticasone&azithromycin
dextromethorphan hbr&guaifenesin
guaifenesin&dextromethorphan hbr

Figure.5 The ‘Upper respiratory tract infectious’ community

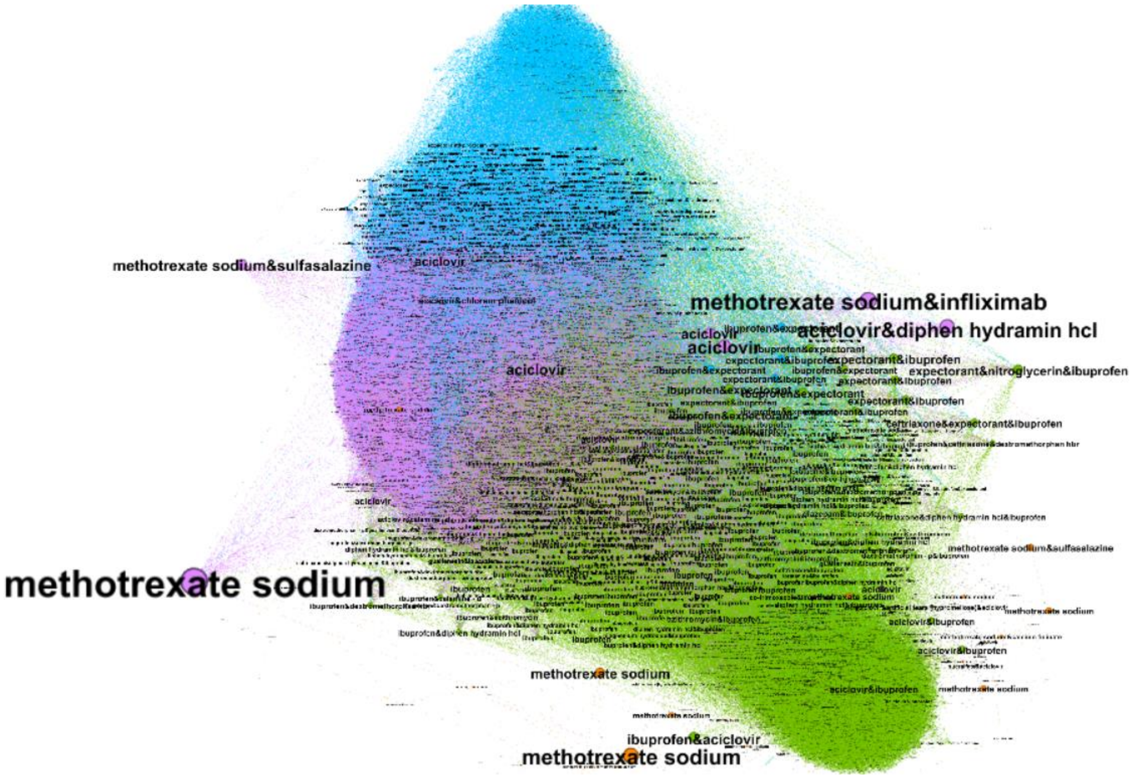
methotrexate sodium&sulfasalazine
aciclovir
methotrexate sodium&infliximab
aciclovir&diphen hydramin hcl
expectorant&ibuprofen
expectorant&nitroglycerin&ibuprofen
ibuprofen&expectorant
ceftriaxone&expectorant&ibuprofen
methotrexate sodium&sulfasalazine
methotrexate sodium
aciclovir&ibuprofen
ibuprofen&aciclovir

Figure.6 The ‘Respiratory and urinary tract infections’ community

Conclusion

Detecting and labeling diseases based on prescriptions presented in this paper. We have used community detection method for identifying groups of medicines that prescribe with each other by physicians. The reason for applying the community detection for this matter was lack of informative attributes in dataset preventing us from using well-known data mining approaches of clustering. The other purpose was the sparsity of data in which led us applying graph mining method instead of a data mining method. Furthermore, we have implemented the Louvain algorithm in this paper for detecting communities. Louvain is a popular greedy algorithm in case of weighted graphs. Due to its fast convergence properties, high modularity, and hierarchical partitioning, Louvain has been widely used in many fields.[45] Our main reason for using this algorithm in this paper was its acceptable time complexity in comparing to other algorithms in this case. Moreover, our validation process consisted of modularity metric. On the other hand, since the modularity varies based on the type of data and context and there is no single right or wrong answer for the community detection process,[46] we have used experts' opinion as an external validation metric for obtaining more accurate results.

The information provided in this paper has many applications in healthcare. One form is by assigning the disease label of prescriptions as claim data, each prescription presents a disease. With this extra information as a data source, scientists can investigate epidemiologic features of society like prevalence and incidence. Moreover, an example of this new data source could be investigating on cost of diseases in society. This information could also be in case of individuals level and measurements based on sex and age category of patients that could affect financial policies in healthcare-related organizations.

References

1. Zhang Y, Qiu M, Member S, Tsai C. Health-CPS : Healthcare Cyber-Physical System Assisted by Cloud and Big Data. 2015:1-8.

2. Suseela BBJ, Jeyakrishnan V. A MULTI-OBJECTIVE HYBRID ACO-PSO OPTIMIZATION ALGORITHM FOR VIRTUAL MACHINE PLACEMENT IN CLOUD COMPUTING. 2014:2319-2322.

3. Herland M, Khoshgoftaar TM, Wald R. Open Access A review of data mining using big data in health informatics. 2014.

4. Manogaran G, Thota C, Lopez D, Vijayakumar V, Abbas KM, Sundarsekar R. Big Data Knowledge System in Healthcare. doi:10.1007/978-3-319-49736-5

5. Denny JC. Chapter 13 : Mining Electronic Health Records in the Genomics Era. 2012;8(12). doi:10.1371/journal.pcbi.1002823

6. Fortunato S. Community detection in graphs. *Phys Rep*. 2010;486(3-5):75-174. doi:10.1016/j.physrep.2009.11.002

7. Newman MEJ. Modularity and community structure in networks. 2006;103(23):8577-8582.

8. Wang M, Wang C, Yu JX, Zhang J. Community Detection in Social Networks : An In-depth Benchmarking Study with a Procedure-Oriented Framework. :998-1009.

9. Li Z, Liu J. A multi-agent genetic algorithm for community detection in complex networks. 2016;449:336-347. doi:10.1016/j.physa.2015.12.126

10. He T, Meng T, Chen L, Deng Z, Cao Z. Parallel Community Detection Based on Distance Dynamics For Large-scale. *IEEE Access*. 2018;PP(c):1. doi:10.1109/ACCESS.2018.2859788

11. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. 2008:1-12. doi:10.1088/1742-5468/2008/10/P10008

12. Dabas C, Nagar H, Kumar G. ScienceDirect ScienceDirect Large Scale Graph Evaluation for Find Communities in Big Data Large Scale Graph Evaluation for Find Communities in Big Data. *Procedia Comput Sci*. 2018;132:263-270. doi:10.1016/j.procs.2018.05.171

13. Chopade P, Zhan J. Community Detection in Large Scale Big Data Networks. (1137443).

14. Chopade P, Zhan J. Networks Using Game-Theoretic Modeling. 2016;XX(X). doi:10.1109/TBDATA.2016.2628725

15. Long H. PT US CR. *Inf Sci (Ny)*. 2018. doi:10.1016/j.ins.2018.03.063

16. Pirouz M, Zhan J. Optimized Label Propagation Community Detection on Big Data Networks. 2018.

17. Ahajjam S, Haddad M El, Badir H. A new scalable leader-community detection approach for community detection in social networks. *Soc Networks*. 2018;54:41-49. doi:10.1016/j.socnet.2017.11.004

18. Karyotis V, Tsitseklis K, Sotiropoulos K. Big Data Clustering via Community Detection and Hyperbolic Network Embedding in IoT Applications. :1-21.

doi:10.3390/s18041205

19. Li X, Cao X, Qiu X, Zhao J, Zheng J. Intelligent Anti-Money Laundering Solution Based upon Novel Community Detection in Massive Transaction Networks on Spark. *Proc - 5th Int Conf Adv Cloud Big Data, CBD 2017*. 2017:176-181. doi:10.1109/CBD.2017.38

20. Bichot, C.-E. and P. Siarry G partitioning. 2013: JW& S. No Title.

21. Hung S, Araujo M, Faloutsos C. Distributed Community Detection on Edge-labeled Graphs using Spark. doi:10.1145/1235

22. Guendouz M, Amine A, Mohamed R. A discrete modified fireworks algorithm for community. 2017:373-385. doi:10.1007/s10489-016-0840-9

23. Moon S, Lee J, Kang M. Scalable Community Detection from Networks by Computing Edge Betweenness on MapReduce. 2014:14-17.

24. Ovelg M. Distributed Community Detection in Web-Scale Networks. 2013:66-73.

25. Saltz M, Prat-pérez A, Dominguez-sal D. Distributed Community Detection with the WCC Metric. :1095-1100.

26. Sharma R, Oliveira S. Community Detection Algorithm for Big Social Networks Using Hybrid Architecture. *Big Data Res*. 2017;10:44-52. doi:10.1016/j.bdr.2017.10.003

27. Ciprian-Octavian Truicˇa, Olivera Novovi´c,Sanja Brdar ANP. No Title. In: *Community Detection in Who-Calls-Whom Social Networks*. Conference: International Conference on Big Data Analytics and Knowledge Discovery; 2018:15.

28. Landon BE, Onnela JP, Keating NL, et al. Using administrative data to identify naturally occurring networks of physicians. *Med Care*. 2013;51(8):715-721. doi:10.1097/MLR.0b013e3182977991

29. Cox M, Ellsworth D. Application-Controlled Demand Paging for Out-of-Core Visualization Page size.

30. Bryant RE, Katz RH, Lazowska ED. Big-Data Computing : Creating revolutionary breakthroughs in commerce , science , and society Motivation : Our Data-Driven World. 2008.

31. Sood SK, Sandhu R, Singla K, Chang V. Sustainable Computing : Informatics and Systems IoT , big data and HPC based smart flood management framework. *Sustain Comput Informatics Syst*. 2018;20:102-117. doi:10.1016/j.suscom.2017.12.001

32. Kulennavar PN. A Survey On Big Data Analytics In Health Care. 2014;5(4):5865-5868.

33. Big data : The next frontier for innovation , competition , and productivity. 2011;(June).

34. Elhoseny M, Abdelaziz A, Salama AS, Riad AM, Muhammad K, Kumar A. A hybrid model of Internet of Things and cloud computing to manage big data in health services applications. *Futur Gener Comput Syst*. 2018. doi:10.1016/j.future.2018.03.005

35. Swan M. THE QUANTIFIED SELF : 2013;1(2):85-99. doi:10.1089/big.2012.0002

36. Krumholz HM. Downloaded from content.healthaffairs.org by Health Affairs on April 5, 2015 at UNIV OF MASSACHUSETTS. 2014. doi:10.1377/hlthaff.2014.0053

37. Taher A, Aboul A, Hassanien E. Dimensionality reduction of medical big data using neural-fuzzy classifier. 2014. doi:10.1007/s00500-014-1327-4

38. Yao Q, Tian Y, Li P, Tian L. Design and Development of a Medical Big Data Processing System Based on Hadoop. 2015. doi:10.1007/s10916-015-0220-8

39. Shaikh AR, Butte AJ, Schully SD, Dalton WS, Khoury J, Hesse BW. Collaborative Biomedicine in the Age of Big Data : The Case of Cancer Corresponding Author : 2014;16:1-5. doi:10.2196/jmir.2496

40. Meyer A, Olshan AF, Green L, et al. Big Data for Population-Based Cancer Research : 2014;75(4):265-269.

41. Szlezák N, Evers M, Wang J, Pérez L. The Role of Big Data and Advanced Analytics in Drug Discovery , Development , and Commercialization. 2014;95(5):3-6. doi:10.1038/clpt.2014.29

42. Udrescu L, Sbârcea L, Topîrceanu A, et al. Clustering drug-drug interaction networks with energy model layouts: Community analysis and drug repurposing. *Sci Rep*. 2016;6(June):1-10. doi:10.1038/srep32745

43. Chautard E, Thierry-mieg N, Ricard-blum S. Interaction networks : From protein functions to drug discovery . A review ´ seaux d ' interactions : de la fonction des prote Les re ` la conception de me ´ dicaments . Une revue a. 2009;57:324-333. doi:10.1016/j.patbio.2008.10.004

44. Brandes U, Borgatti SP, Freeman LC. Maintaining the duality of closeness and betweenness centrality ଝ. *Soc Networks*. 2016;44:153-159. doi:10.1016/j.socnet.2015.08.003

45. Que X, Checconi F, Gunnels JA. Scalable Community Detection with the Louvain Algorithm. 2015. doi:10.1109/IPDPS.2015.59

46. Mastering Spark for Data Science - Andrew Morgan, Antoine Amend, David George, Matthew Hallett - Google Books.